Interrater reliability of research assistants performing data collection for the Ohio Perinatal Research Network

Joan P Randle, Senior Nursing Student

Dr. Thelma Patrick, PhD, RN, Faculty Advisor

The Ohio State University

College of Nursing

Abstract

The purpose of this study was to investigate the interrater reliability of data collected through chart abstraction by research assistants affiliated with The Perinatal Research Repository (PRR), a division of the Ohio Perinatal Research Network (OPRN). Assessing interrater reliability is an important component of chart abstraction to ensure the precision of data collected. The Perinatal Research Repository collects data from premature infants, their mothers and fathers for current and future research. Twenty charts previously abstracted by the senior research assistant were randomly selected for this study. After a second abstraction the identified variables were compared for reliability. A percentage of agreement was determined. If the interrater reliability met the expected standard, the study will be terminated after ten charts are reviewed. When interrater reliability did not meet the established standard, the specific variations in the procedures of carrying out the abstraction and interpreting the results was identified and the OPRN Maternal Abstraction Form was modified. The Perinatal Research Repository is an important resource for social context data, clinical data, demographic data, environmental context data, and biospecimen data related to discovering and implementing actions that measurably reduce prematurity. Reliable data abstraction ensures the quality of research is preserved.

Interrater reliability of research assistants performing data collection for the Ohio Perinatal Research Network

Introduction

The Institute of Medicine reports that over half a million babies are born preterm annually, with an associated health care cost of over 26 billion dollars (NIM, 2010). The rate of premature birth nationally is approximately 12.3%, and in Ohio is 12.6% according to the March of Dimes 2011 Premature Birth Report Card. In Franklin County the preterm birth rate is 13% and as high as 20% among high risk populations. These conditions have significant impact on mortality and morbidity. (https://www.aamc.org/newsroom/reporter/march2012/276796/innovations.html).

The study of factors that contributes to pregnancy complications and prematurity is an important endeavor. The March of Dimes reports that half of spontaneous preterm labors and 40% of premature births have an unknown etiology. Researchers have hypothesized that genetics, environmental factors, and maternal infection are involved in the process of premature delivery. There are difficulties in investigating factors associated with and consequences of prematurity, especially since premature birth is not usually planned or expected. Even when women are enrolled in a study at the time of diagnosis of a maternal complication leading to prematurity or following an adverse event prompting an emergent early delivery, much of the data needed must be obtained from recall.

Large data repositories can be as a source of data, and sometimes biosamples, to perform hypothesis generating studies that will advance the study of a less frequently occurring disorder.

These large databases provide a predetermined set of data, and investigators are then able to utilize the large number of potential study participants for epidemiological research (Pass, 2010). Data are collected from different sources, depending upon the goal for the database.

The Database for Preterm Birth (dbPTB) is an example of a web-based aggregation tool whose focus is investigating genetic variations and pathways involved in preterm birth. This database is unique in that it integrates the ability to gather data from published sources and genetic samples (http://ptbdb.cs.brown.edu/dbPTBv1.php) (Uzun, et al, 2012). The Vermont Oxford Network Quality Collaborative (VON), established in 1988, is another example of a large data repository whose primary mission is to gather and maintain data on very low birth weight babies (http://www.vtoxford.org). There are over 900 member Neonatal Intensive Care Units from around the world that contribute to and access this database for research purposes.

The use of large pre-existing databases for research purposes is efficient, cost effective, and lends itself well to a variety of research methods (Mann, 2003). As electronic charting is adopted more widely in healthcare, the possibility of collecting and managing large quantities of data will enable investigators to access a vast array of information quickly. Cohort studies, cross sectional studies and case control studies are examples of types of research that can be developed from pre-existing databases. For example, study participants for a retrospective case control study can be selected based on outcomes of interest from large databases (Mann, 2003). When the conditions and/or outcomes being studied are rare, as in the case with prematurity, this type of study can be very beneficial (Mann, 2003).

The Ohio Perinatal Research Network (OPRN) was established with the mission of improving “the lives of children and families in Ohio and beyond by discovering and implementing actions that measurably reduce prematurity associated morbidity and mortality”

(http://www.nationwidechildrens.org/ohio-perinatal-research-network). The Ohio Perinatal Research Network supports the NCH Section of Neonatal Medicine which has affiliated with The Neonatal Research Network, developed by the National Institute of Child Health and Development (NICHD) whose mission it is to improve the care and outcome of neonates.

The Ohio Perinatal Research Network has a different focus, collecting pertinent information from parents and infants as well as biological specimens. The Perinatal Research Repository is a data warehouse affiliated with the Ohio Perinatal Research Network. Clinically relevant information and biological specimens from premature infants, their mothers and fathers are collected for use in research related to the diagnosis, prevention and treatment of preterm birth and diseases and complications related to preterm birth. This registry provides network investigators with information regarding important variables that affect the short and long-term health outcomes of premature infants.

The purpose of this investigation is to describe methods for assessing reliability of data abstraction among research staff, and to conduct such an assessment for a specific number of variables in the database. Following this assessment, we will identify areas where the expected percent agreement has not been achieved, reach consensus about necessary revisions to improve the abstracted data, and revise definitions for abstraction as needed. This information will benefit the quality and reliability of data in the OPRN repository.

Review of Literature

A search of PubMed was conducted in order to identify studies that investigated the reliability and validity of data collection among trained research assistants, provided examples of studies that could serve as exemplar studies to better understand the process and utilization of data from large databases, and that search was limited to examples of maternal complications or prematurity.

High caliber data is the foundation of reputable medical registries. Arts (2002) defines "good quality data" as "the totality of features and characteristics of a data set that bear on its ability to satisfy the needs that result from the intended use of the data." In order for data to fulfill this definition, the data must be both complete and accurate (Pass, 2010). Complete and accurate data collection is, however, dependent on the development of a strong clinical research form (CRF) onto which data is initially coded (Pass, 2010). Errors tend to increase in relation to the number of times data are manipulated, or transferred, from chart to CRF to final database, as well as when terms are not well defined and instructions are unclear (Pass, 2010).

For research projects dependent upon data repositories to provide meaningful information, collection and recording of clinically relevant data must be accurate. Reliability and validity of data collection can be enhanced with consistent definitions and clearly written directions for abstractors of varying medical literacy levels. Without consistency the data become falsely skewed and unreliable, potentially impacting patient care.

Individual reviewer errors include data misinterpretation, errors originating in the original patient document, and random errors of data entry (Goldberg, 2008). Other data errors may

occur which are intrinsic to the CRF or due to ambiguous and/or inconsistently applied definitions. The process of collecting clinically relevant data from patient charts and transferring the information from chart to CRF to permanent database is tedious and time consuming, accounting for a reported discrepancy rate of 13.5% to 27% (Goldberg, 2008). High human error can be a result of poor CRF design, where data fields do not flow intuitively from a cognitive and/or a visual perspective, or due to misinterpretation of data (Goldberg, 2008).

Interrater reliability studies are useful when retrospective data collection from patient charts is a regular part of clinical research. Chart abstraction, gleaning information from patient records, affords an easy, non-invasive method of data retrieval. Reliability among abstractors, however, is paramount to ensure that data is accurately and consistently recorded. Inconsistencies may be the result of unclear data collection questions or due to abstractors' differing interpretations of data. Interrater reliability audits can identify areas that are prone to subjective interpretation by different abstractors or ambiguous terminology in the abstracting form itself.

Green and Lewis (1986) suggest that internal consistency standards be set and calculated before data collection begins. Establishment of interrater reliability standards and subsequent regular audits of abstractors' coding ensure the reliability of data collected. Interrater reliability "estimates the amount of error in an instrument's score that is caused by the observation, rating or coding process;" or rather, the extent of variation due to individual understanding, perception, or rater fatigue (Green & Lewis, 1986). Interrater reliability can be calculated using either a kappa coefficient or a percentage. The kappa coefficient is a useful tool in calculating interrater reliability as it adjusts for chance agreement among raters. Percent agreement is also used to calculate interrater reliability.

An exemplar case-control study of prematurity, *Fetal Infants: The Fate of 4172 Infants With Birth Weights of 401-500 Grams - The Vermont Oxford Network Experience (1996-2000)* illustrates the necessity of uniform definitions when data are collected from multiple sources (Lucey, et al, 2004). The cohort of infants studied was developed from the Vermont Oxford Network Quality Collaborative base, initially selecting 4172 infants from 346 participating institutions who met the primary inclusion criteria of birth weight. The study then analyzed specific characteristics and outcomes of interest. As this study was interested in all infants born between 1996-2000 within the birth weight parameters, mortality was not exclusionary. Delivery Room interventions, population characteristics of both surviving and deceased infants were compared and outcomes of survivors identified.

The reliability of data collected from the 346 institutions is high owing to the use of the VON's standardized manual of operations which included accepted terminology usages and definitions. In addition, each institution agreed to self-administered internal audits of data collection to verify accuracy and completeness. The first tier research questions investigated pertained to delivery room interventions and infant survival beyond delivery room. Second tier research questions filtered surviving infants into one of three sub-groups: those that died while in the NICU; those that survived to discharge; and those with an unknown survival status. Specific population characteristics were then collected for delivery room survivors versus delivery room deaths, and for NICU survivors versus NICU deaths. Lastly, morbidities and outcomes from the NICU survivors group were evaluated.

Identified strengths of this case-controlled study include the large sample size and the agreement of participating institutions to adhere to uniform definition. Limitations, on the other hand, include the lack of information regarding decision making in delivery room (the value

judgements used by physicians as to the viability of the infant); and use of birth weight stratification as the tool for reporting infant morbidities (instead of the more common practice of using gestational age) (Lucey, et al, 2004).

Clearly, the valid and reliable data collected from various hospitals is the foundation upon which this study and the Vermont Oxford Network rely. The foundation of the data itself is the initial collection efforts of each institution, and their compliance with the VON's directive for each institution to perform verification audits. The use of VON's Manual of Operations also gives each institution a resource for reinforcement of uniform definitions.

A second study, entitled *Plasma uric acid remains a marker of poor outcome in hypertensive pregnancy: a retrospective cohort study,* (Hawkins, T., et al, 2012) is another exemplary article which is illustrative of a retrospective cohort and nested case-control study. This study investigated a cohort of pregnant women who were referred for management of preeclampsia or gestational hypertension (n = 1880). The database used was developed from two separate databases and represents women who were either identified as hypertensive by obstetrical staff and those who were referred for management by independent renal physicians.

Accepted guidelines delineating acceptable definitions for "preeclampsia" and "gestational hypertension" were developed to insure that the classisfication of women into the two disease cohorts was accurate. Laboratory values and subsequent changes of same were collected and relationships identified with the intention of determining if, and to what extent, hyperuricemia and hemoconcentration were implicated in adverse outcomes for women who are have hypertension during pregnancy, as well as poor infant outcomes (Hawking, T, et al).

After determining the positive relationship between hyperuricemia and adverse maternal/fetal outcome, a nested case-control study was conducted. This nested case-control

study determined women who were experiencing gestational hypertension with hyperuricemia were at increased risk of preterm birth and infants were at increased risk of having substandard birth weights.

Collecting relevant laboratory values from patient records demands sustained attention to detail and dedication to the process of deconstructing patient records in order to fully comply with data collection requirements. These standards are vital to completing clinical research forms as thoroughly as possible.

These exemplars provided positive examples of the techniques that must be utilized so that investigators have confidence in the data. Without this confidence, patient classification into study groups, specific factors under study, and outcomes of interest could all be suspect.

Methods

The purpose of this study was to assess interrater reliability of chart abstractions of specific and defined data from The Perinatal Research Repository. The Perinatal Research Repository is an important resource for social context data, clinical data, demographic data, environmental context data, and biospecimen data related to discovering and implementing actions that measurably reduce prematurity. Reliable data abstraction ensures the quality of research is preserved.

Several of the steps for ensuring the veracity of data were included in the protocol for review. A Manual of Operations was developed by clinical and research experts to provide definition and clarification of questionable or ambiguous items on the abstraction form. Number of pregnancies, for example, is clarified as actual number of pregnancies, specifying that multiple gestation pregnancies are documented as one pregnancy. Reviewers were given the Manual of Operations to familiarize themselves with accepted definitions.

Consistency of abstraction results was calculated to a preset percentage of total criteria for which there was agreement in relation to total number of criteria reviewed. The target goal of 90% for interrater reliability was established.

Using criteria developed to describe the perinatal experience of each mother and infant, Reviewer #1, a study staff member who is assigned routine chart abstraction, performed the first abstraction. This investigator (Reviewer #2) participated in the process of chart abstraction for

a subset of charts that had already been abstracted by that study staff member. Reviewer #2, a student, selected 20 charts at random and performed an independent duplicate abstraction. Abstraction results from Reviewers #1 and #2 were compared and interrater reliability was calculated. Reviewer #3, an investigator for the study, selected two charts at random from the 20 charts abstracted by Reviewers # 1 and #2. These results were compared with the findings of reviewers #1 and #2.

After all abstractions were completed, the reviewers met to compare the abstractions and analyze the discrepancies. Each of the 20 charts was analyzed in detail, examining each item for which there was disagreement. For items with a high percentage of disagreement, the Manual of Operations was consulted for clarification. Modifications to the Manual of Operations were made if the team determined the manual's definition or instruction was easily misinterpreted as evidenced by disagreement among the team.

Research Results

Review of chart abstraction data was completed by the identified reviewers and compared. Interrater reliability was calculated with the number of total criteria review as denominator, and the number of criteria for which there was agreement as the numerator. Interrater reliability for total chart abstractions ranged from 86% to 94%. Reviewer #1 and #2 achieved an interrater reliability of 89% while reviewers #2 and #3 achieved an interrater reliability of 95%. Reviewers #1 and 3 had an interrater reliability score of 88%.

Reviewers #1, 2, and 3 met to discuss discrepancies. Variables for which the reviewers' abstractions were discrepant included: date of the first prenatal visit, race of the baby, date of the first ultrasound, sonogram at >24 weeks, antibiotics administered to the mother, and last documented weight. These variables may have had more than one potential response in the chart, and instructions as to what response was the optimal to use was not defined. Other issues discovered were baby's race, an item that was also difficult to find documented in the paper chart; or more likely, was assumed to be based on documented maternal race. The administration of antibiotics to the mother was another area that was difficult to agree upon, as there were at least two items which asked about said administration. One of these items was specific to antibiotics for urinary tract infection, which the other item inquired about antibiotic use for various conditions such as Group B Strep.

Another factor that contributed to discrepancies was the expectation that the term "missing" information or information "not recorded" were defined in similar terms, but when

utilized, the reviewer chose which of these terms would be recorded. This resulted in different terms being recorded for the same absence of information in the chart. Finally, there were discrepancies in recorded data due to the reviewer's use of two different versions of the abstraction form. These different versions were the result of modifications had been made prior the abstraction by Reviewer #2, which the primary abstraction had been recorded on the earlier version.

Once discrepancies were resolved, interrater reliability for all three reviewers reached 94%, exceeding the preset standard. Modifications to the abstraction form or Manual of Operations were made based on the identified concerns.

Discussion

Access to, and review of, maternal and infant records is imperative to identifying maternal and infant risk factors resulting in prematurity. Obtaining accurate measures related to the variables that affect the short and long-term health outcomes of premature infants and their families is the goal of chart abstraction in the Center. The generation of quality descriptive and nested case-control that can lead to a better understanding of contributing factors and pathogenesis depends on the quality of the available data (Green & Lewis, 1986; Hennekens & Buring, 1987; Mann, 2010).

Interrater reliability was increased after meeting with reviewers to discuss reasons for the discrepancies, achieving the target 90% reliability rating. Issues that increase the likelihood of discrepant findings can be individual or process oriented. For example, reviewer fatigue was a prominent issue that was revealed. Green and Lewis (1986) suggest that as reviewer interest in the project declines, error rate, in the form of observational or documentation errors increase. Reviewer fatigue is also more prone to occur when the reviewer becomes very familiar with the CRF, the routine of the task, or is distracted. Quality control measures such as random audits can reduce the incidence of rater error (Green & Lewis, 1986). Limiting the number of hours per day or number of charts per day that are abstracted may also reduce reviewer fatigue. Developing an abstraction instrument which is reflective of the information as it is recorded in the charting system (paper or electronic) would streamline the search for relevant data in said charts, as well as help ensure consistency in documentation among reviewers.

Establishment of a regular audit schedule will decrease the incidence of error and maintain the integrity of the data base. Regular review of the Manual of Operations concurrent

with audits will also refresh reviewers' understanding of terminology. A record of all revisions to the abstraction form or definitions, including the date of the change and the records to which that modification applies is paramount to informing investigators who plan to use the data of potential conflicts.

Summary

In this study, we conducted an assessment of interrater reliability of a specific dataset for retrospective study. Consistency and precision in the chart abstraction process are critical to retrospective studies relying on chart data in the study of maternal and infant factors that contribute to prematurity and related complications. The Perinatal Research Repository is an important resource for social context data, clinical data, demographic data, environmental context data, and biospecimen data related to discovering and implementing actions that measurably reduce prematurity. Reliable data abstraction ensures the quality of research is preserved.

References

Arts D. G., de Keizer N. F., Scheffer G. J. (2002). Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc 2002; 9:600–611.

Baraldi, E., Carraro, S., Filippone, M. (2009) Bronchopulmonary Dysplasia: Definitions and long term respiratory outcome. Early Human Development; 85 (2009) S1-S3.

Been, J., Rours, I., Kornelisse, R., Jonkers, F., deKrijger, R., Zimmermann, L. (2010). Chorioamnionitis alters the response to sufactant in preterm infants. JPeds, 10.1016/.jpeds.2009.07.044.

Carter, B.M., Holditch-Davis, D. (2008) Risk Factors for NEC in preterm infants: how race, gender and health status contribute. Published in final edited form as: Adv Neonatal Care. 2008 October; 8(5): 285–290. doi: 10.1097/01.ANC.0000338019.56405.29

Chess, P., D'Angio, C., Pryhuber, G., Maniscalco, W. (2006) Pathogenesis of bronchopulmonary dysplasia. Seminars in Perinatology; 10.1053/j.semperi.2006.03.03.

Cruz, C., et al, (2009). Interrater reliability and accuracy of clinicians and trained research assistants performing prospective data collection in emergency department patients with potential acute coronary syndrome. Ann Emerg Med, doi10.1016j.annemergmed.2008.11.0235.

Ehrenkranz, RA, Walsh, MC, Vohr, BR, Jobe, AH, Wright, LL, Fanaroff, AA, Wrage, LA, Poole, K. Validation of the National Institutes of Health consensus definition of bronchopulmonary dysplasia. Pediatrics 2005; 116: 1353-1360.

Farstad, T., Bratlid, D., Medbo, S., Markestad, T., (2011). Bronchopulmonary dysplasia-prevalence, severity and predictive factors in a national cohort of extremely premature infants. ActaPaedeatrica 2011; 100; pp53-58.

Gephart, S., McGrath, J., Effken, J., Halpern, M. (2012). Necrotizing enterocolitis risk. Advances in Neonatal Care; Vol 12, No 2, pp 77-87.

Goldberg, S., Neimierko, A., Turchin, A. (2008) Analysis of Data Errors in Clinical Research Databases ; AMIA Annu Symp Proc. 2008; 2008: 242–246. PMCID: PMC2656002

Hawkins, T., Roberts, J., Mangos, G., Davis, G., Roberts, L., Brown, M. (2012) Plasma uric acid remains a marker of poor outcome in hypertensive pregnancy: A retrospective cohort study. BJOG 2012;119:484-492.

Jadcherla SR, Peng J, Moore R, Saavedra J, Shepherd E, Fernandez S, Erdman SH, Di Lorenzo C. 2011. Impact of personalized feeding program in 100 NICU infants: A novel pathophysiology-based approach for better outcomes. J Pediatr Gastroenterol Nutr. (Formally Accepted)

Jadcherla SR, Wang M, Vijayapal AS, Leuthner SR. 2010. Impact of prematurity and co-morbidities on feeding milestones in neonates: a retrospective study. J Perinatol. 2010 Mar;30(3):201-8.. Vol. 30 (3), no. J Perinatol. 2010 Mar;30(3):201-8. (March): 201-208.

James, L., Demaree, R., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology,69(1), 89-98.6.

Jobe, A. (2006) The New BPD. NeoReviews; Vol 7, No 10: October 2006.

Kennedy, K., Tyson, J., Chamnanvanikji, S. (2000). Early versus delayed initiation of progressive enteral feedings for parenterally-fed low birth weight or preterm infants. Cochrane Database Syst Review. 2000; (2):CD001970.

Lucey, J.F., Rowan, C.A., Shiono, P., Wilkinson, A.R., Kilpatrick, S., Payne, N. R., Horbar, J., Carpenter, J., Rogowski, J., Soll, R.F. (2004). Fetal Infants: The fate of 4172 infants with birth weights of 401-500 grams - The Vermont Oxford Network experience (1996-2000). Pediatrics, 2004; 113; 1559

retrieved from http://pediatrics.aapublications.org/content113/6/1559.full.html

Mann, C. J. (2003). Observational research methods. Research design II: Cohort, cross sectional, and case-control studies. Emerg Med J 2003; 20: 54-60, doi:10.2236/emj.20.1.54.

Monsen, K., Lytton, A., Ferrari, S., Haler, K., Radosevich Kerr, M., Mitchell, S., Brandt, J. (October 2011). Evaluating reliability of assessments in nursing documentation. Online Journal of Nursing Informatics, 15(3), Retrieved from http://ojni.org/issues/?p=899

Pass, H. (2010). Medical Registries: Continued Attempts for Robust Quality Data. Journal of Thoracic Oncology: June 2010 - Volume 5 - Issue 6 - pp S198-S199

doi: 10.1097JTO.0b013e3181dcf957

Poore, M., Barlow, S., Wang, J., Estep, M., Lee, J. (2008). Respiratory treatment history predicts suck pattern stability in preterm infants. J Neonatal Nurs. 2008; 14 (6): 185-192. doi: 10.1016/j.jnn.2008.07.006.

Pan, L., Ferguson, D., Schweitzer, I., Hebert, P. (2005). Ensuring high accuracy of data abstracted from patient charts: the use of a standardized medical record as a training tool. Journal of Clinical Epidemiology 58, 918-923.

Samara, M., Johnson, S., Lamberts, K., Marlow, N., Wolke, D. (2009). Eating problems at age 6 years in a whole population sample of extremely preterm children. Developmental Medicine and Child Neurology 10.11111/j.1469-8749-2009.03512.x.

Uzun A, Laliberte A, Parker J, Andrew C, Winterrowd E, Sharma S, Istrail S, Padbury JF. (2012) dbPTB: a database for preterm birth. Database (Oxford). 2012 Feb 8;2012:bar069.

Uzun, A., Surendra, S., Padbury, J. (2012) A Bioinformatics Approach to Preterm Birth Am J Reprod Immunol. 2012 April; 67(4): 273–277. Published online 2012 March 5. doi: 10.1111/j.1600-0897.2012.01122.x

Web Resources

http://www.managedcaremag.com/archives/1001/1001.preterm.html

http://www.marchofdimes.com/news/nov01_2011.html